Lecture 4: Probability and Statistics

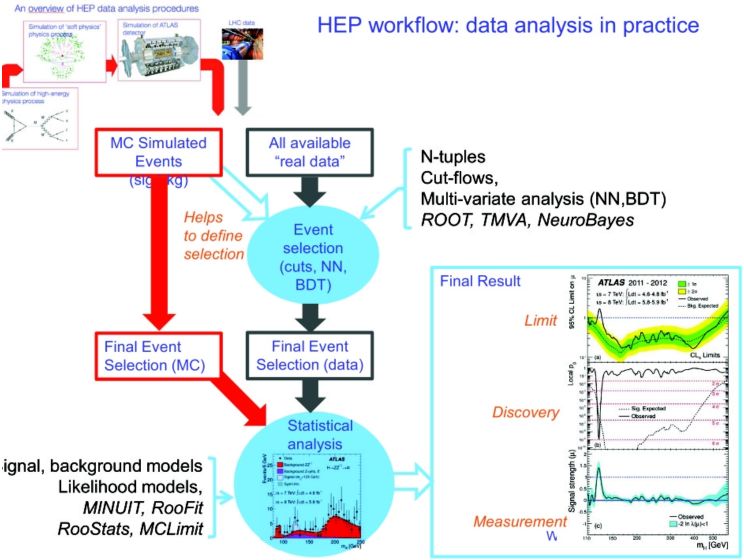
September 6, 2016

Introduction

- Physics is based on experimental measurements
- Must understand precision and accuracy of these measurements
- Must also determine whether data is consistent with our theory and whether new physics could be hiding in the data

Statistics provides the tools to do this

How particle physicists analyze data



W. Verkerke, 2014 European Particle Physics Summer School

Probability: Basic Definitions and Axioms

- Probability P is a real-valued function defined by axioms:
  1. For every subset A in S, $P(A) > 0$
  2. For disjoint subsets ($A \cap B = 0$), $P(A \cup B) = P(A) + P(B)$
  3. $P(S) = 1$
- Bayes Theorem:
  (Conditional Probability $P(A|B) \equiv$ prob of A given B)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- Law of Total Probability

$$P(B) = \sum_i P(B|A)P(A_i)$$

- Together these give:

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_i P(B|A_i)P(A_i)}$$

Probability: Random variables and PDFs

- For continuous variable x, probability density function (pdf):
  - $f(x;\theta) \equiv$ prob that x lies between x and $x+dx$
  - θ represents one or more parameters
    Won't always carry θ along
- Cumulative probability

$$F(a) = \int_{-\infty}^{a} f(x)dx$$

  Probability that $x < a$.
- For discrete variables, replace integral with sum
- For any function $u(x)$, expectation value:

$$E[u(x)] \equiv \langle u(x) \rangle = \int_{-\infty}^{\infty} u(x) f(x) dx$$

PDF Moments: Mean and Variance

- Mean value:

$$\mu \equiv \int_{-\infty}^{\infty} x f(x) dx$$

- Variance:

$$\sigma^2 \equiv Var(x) = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2$$

σ is called the "standard deviation."

These basic definitions are used essentially everywhere. If we know the pdf, we know how to determine the mean and σ

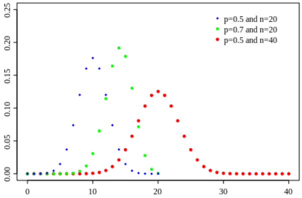
Binomial Distribution [Discrete]

- Random process with two possible outcomes
- p =Prob of outcome #1, $q = 1 - p$ =Prob of outcome #2
- In n trials prob of getting outcome #1 exactly k times is

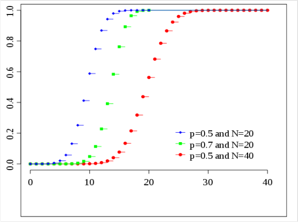
$$f(k; n, p) = \binom{n}{k} p^k q^{n-k} \quad \text{where} \quad \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- $\mu = np$; $\sigma^2 = npq$

Binomial PDF



Binomial Cumulative PDF



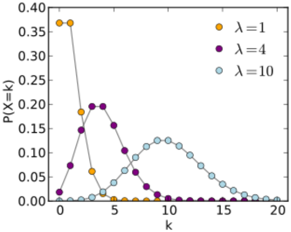
Poisson Distribution [Discrete]

- Prob of finding exactly k events in the interval between x and $x + dx$ if the events occur with an average rate in that interval of λ.

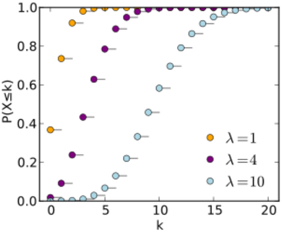
$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- $\mu = \lambda$; $\sigma^2 = \lambda$
- For large λ, approaches a Gaussian

Poisson PDF



Poisson Cumulative PDF



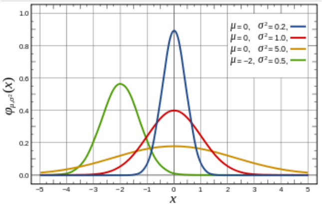
Normal (Gaussian) Distribution [Continuous]

Theorem (Central Limit Theorem)

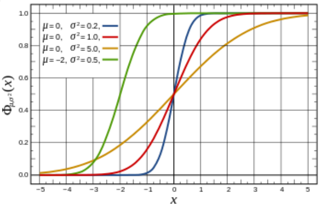
Given random sample $(x_1, x_2, ...x_n)$ drawn from pdf with mean μ and variance σ, if mean is $S/n = 1/n \sum_1^n x_i$, distribution of S/n approaches normal distribution as $n \to \infty$ independent of pdf

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Gaussian PDF



Gaussian Cumulative PDF

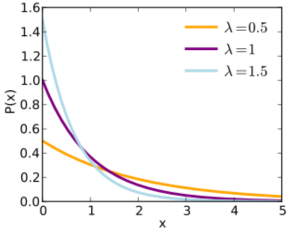


Exponential Distribution [Continuous]

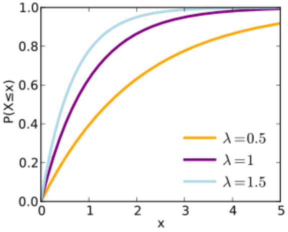
Number of events lost per unit length proportional to number of events

$$
\begin{aligned}
f(x;\lambda) &= \lambda e^{-\lambda x} \\
\mu &= \frac{1}{\lambda};\ \ \sigma^2 = \frac{1}{\lambda^2}
\end{aligned}
$$

Exponential PDF



Exponential Cumulative PDF



Statistical Estimators

- One aim of statistical analysis: estimate true value of one or more parameters from experimental data and understand the uncertainty on that measurement
- Important characteristics a good estimator are:
  - Consistency: If amount of data large, estimate converges to true value
    - Bias: Difference between expectation value of estimator and true value of parameter
  - Robustness: Estimator doesn't change much if true pdf differs from assumed pdf (eg tails in distributions)
- We also want to know the uncertainty on our estimate (how far might the true parameter be from our estimate due to statistical fluctuations in the ensemble of measurements)

Likelihood Function

- Likelihood $\mathcal{L}(x;\theta)$ is probability that a measurement of x will yield a specific value for a given theory
  - To determine likelihood, must know both the theory and the values of any parameters the theory depends on
- If we have an ensemble of measurements, overall likelihood obtained from product of the likelihoods for the measurements

$$\mathcal{L}(x;\theta) = \prod_{i=1}^{n} \mathcal{L}_i$$

Here θ can represent one or more parameters

Log Likelihood

- To estimate parameter(s) θ, maximize the likelihood
- Usual technique to find maximum, set derivative equal to zero
- Easier to maximize than $\ln \mathcal{L}$

$$\begin{aligned}
\frac{\partial \ln \mathcal{L}}{\partial \theta} &= \frac{\partial}{\partial \theta} \ln \prod_{i=1}^{n} \mathcal{L}_i \\
&= \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \ln \mathcal{L}_i \\
&= 0
\end{aligned}$$

- If several θ_i can minimize with respect to each
  - We'll come back to correlations in a few minutes

Poisson example of likelihood

- N independent trials with results n_i
- Likelihood function for observing n_i if true mean is μ

$$\mathcal{L}(n_i;\mu) = \frac{e^{-\mu}(\mu)_i^n}{n_i!}$$

Product over N measurements:

$$\begin{aligned}
\mathcal{L}(data;\mu) &= \prod_{i=1}^{N} \frac{e^{-\mu}(\mu)^{n_i}}{n_i!} \\
\ln \mathcal{L} &= \sum_i \left(-\mu + n_i \ln \mu - \ln(n_i!)\right) \\
&= -N\mu + \left(\sum_i n_i\right) \ln \mu + constant \\
\frac{\partial \ln \mathcal{L}}{d\mu}|_{\hat{\mu}=\mu} &= -N + \frac{\sum_i n_i}{\mu} = 0 \\
\hat{\mu} &= \frac{1}{N} \sum_{i=1}^{N} n_i
\end{aligned}$$

As expected, the best estimator is the mean value

Gaussian example of likelihood

$$G(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Now take derivative of the log likelihood:

$$\begin{aligned}
\frac{\partial}{d\mu}(\ln \mathcal{L})\,|_{\hat{\mu}=\mu} &= \frac{\partial}{d\mu}\left(-\sum_i \frac{(x_i-\mu)^2}{2\sigma^2} + const\right) \\
&= -\sum_i \frac{(x_i-\mu)}{\sigma^2}|_{\mu=\hat{\mu}} = 0 \\
\Rightarrow \quad \hat{\mu} &= \frac{1}{N}\sum_i x_i
\end{aligned}$$

- Warning: The unbiased estimator for σ is

$$\hat{\sigma} = \frac{1}{N-1}\sum_i (x_i-\mu)^2$$

I won't bother to prove this!

Binned vs unbinned likelihood functions

- Likelihood formalism works for any well behaved probability density function
- The product of the likelihood is a product over measurements
- We can define what we mean by a measurement
- Example: Measure the lifetime of particle of a given species from an ensemble of such particles produced at time $t = 0$ that decay at time t:

$$f(t) = \frac{1}{\tau} e^{-t/\tau}$$

  Two ways to construct a likelihood:

  1. For each decay i measure t_i and take the product of all measured times to get $\mathcal{L}$ (unbinned likelihood)
  2. Make a histogram of the number of decays in bins of time. Now, the measurement is the number of decays in each bin i (binned likelihood)

You will have a chance to try this in practice on problem set # 3

Connecting the Log Likelihood to the χ^2

- From previous page, for Gaussian case

$$\ln \mathcal{L} = -\sum_i \frac{(x_i - \mu)^2}{2\sigma^2} + const$$

- Compare this to

$$\chi^2 \equiv \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^2}$$

- By inspection, for the case of a Gaussian distribtuion

$$\chi^2 = -2 \ln \mathcal{L}$$

- Note: The likelihood formulation works for all pdf's and is therefore more general!

The Method of Least Squares

- Assume our measurements are made with high enough statistics that we can assume we are in the Gaussian regime
- We want to find the best estimates of the parameters of function that describes the data
- Do this by minimizing the scatter of data from fit function, taking into account uncertainties on data points
- Scatter defined in terms of χ^2:

$$\chi^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^2}$$

- We can write the χ^2 in terms of our observables

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - F(x_i, \theta))^2}{\sigma_i^2}$$

- Minimize χ^2 with respect to θ (or multiple θ_i)
- Useful in case of high statistics samples where minimizing $-\ln \mathcal{L}$ slow

Correlated Variables

- Often variables we fit for are not independent
- When doing minimization, correlations must be taken into account
- Reminder: variance is:

$$\sigma^2 \equiv Var(x) = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2$$

- Define covariance $Cov[x, y]$ as

$$cov[x, y] == \int_{-\infty}^{\infty} xyf(x, y) dxdy - \mu_x \mu_y$$

- If x and y are uncorrelated, independent variables, then

$$cov[x, y] = 0 \ \text{ for } x \neq y$$

The covariance matrix (Gaussian example)

- If x and y are independent variables

$$
\begin{aligned}
G(x,y|\mu_x,\sigma_x,\mu_y,\sigma_y) &= \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}} \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{(y-\mu_y)^2}{2\sigma_y^2}} \\
\frac{\partial^2}{d\mu_x^2}(\ln\mathcal{L}) &= -\sum_i \frac{1}{\sigma_x^2}
\end{aligned}
$$

  Second derivative wrt μ proportional to $\frac{1}{\sigma^2}$

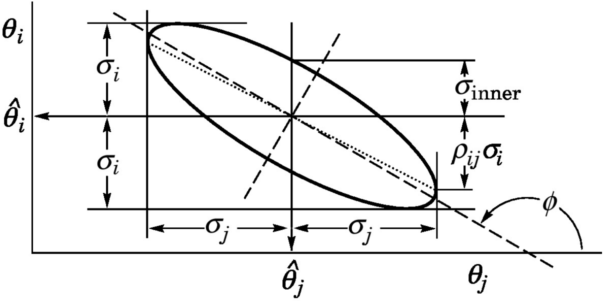
- Now remove assumption that x and y are uncorrelated
- Covariance matrix defined by

$$
\left\langle \hat{V}^{-1} \right\rangle_{ij} = -\frac{\partial^2 \ln\mathcal{L}}{\partial\mu_i \partial\mu_j}
$$

- For binned likelihood in region of large N, where likelihood can be reduced to a χ^2

$$
\left\langle \hat{V}^{-1} \right\rangle = \frac{1}{2}\frac{\partial^2\chi^2}{\partial\mu_i\partial\mu_j}
$$

Effect of Correlated Uncertainties



- Standard error ellipse for two parameters with a negative correlation
- Slope related to correlation coefficient $d\theta_i/d\theta_j$
  - The θ parameters here correspond to the μ parameters on the previous page
- Correlation matrix typically determined from data numerically during fitting procedure

Propagation of Errors

- Good description found on wikipedia:
  http://en.wikipedia.org/wiki/Propagation_of_uncertainty
- Basic expression is

$$\sigma_f^2 = \left(\frac{\partial f}{\partial \alpha}\right)^2 + \left(\frac{\partial f}{\partial \beta}\right)^2 + 2\frac{\partial f}{\partial \alpha}\frac{\partial f}{\partial \beta}COV_{\alpha\beta}$$

  for case where our model has two parameters α and β
- Extension to more dimensions usually expressed as a matrix
- In case of uncorrelated parameters, reduces to the usual expression you saw in undergrad lab

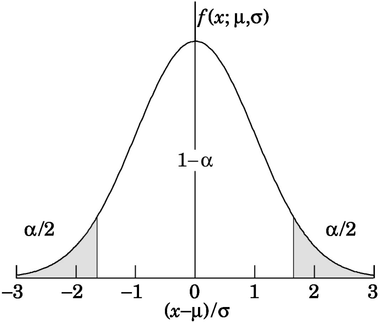
Confidence Intervals

- Using frequentist language: fraction of result is not between x_ℓ and x_u is

$$1 - \alpha = \int_{x_\ell}^{x_u} P(x;\theta)dx$$

  Warning: some authors call this α rather than $1 - \alpha$

- Example for a Gaussian distribution



Confidence Levels for Two Common Distributions

- Gaussian

Table 38.1: Area of the tails α outside $\pm\delta$ from the mean of a Gaussian distribution.

α	δ	α	δ
0.3173	1σ	0.2	1.28σ
4.55×10^{-2}	2σ	0.1	1.64σ
2.7×10^{-3}	3σ	0.05	1.96σ
6.3×10^{-5}	4σ	0.01	2.58σ
5.7×10^{-7}	5σ	0.001	3.29σ
2.0×10^{-9}	6σ	10^{-4}	3.89σ

- Poisson

Table 38.3: Lower and upper (one-sided) limits for the mean μ of a Poisson variable given n observed events in the absence of background, for confidence levels of 90% and 95%.

	$1-\alpha$ =90%		$1-\alpha$ =95%	
n	$\mu_{\rm lo}$	$\mu_{\rm up}$	$\mu_{\rm lo}$	$\mu_{\rm up}$
0	–	2.30	–	3.00
1	0.105	3.89	0.051	4.74
2	0.532	5.32	0.355	6.30
3	1.10	6.68	0.818	7.75
4	1.74	7.99	1.37	9.15
5	2.43	9.27	1.97	10.51
6	3.15	10.53	2.61	11.84
7	3.89	11.77	3.29	13.15
8	4.66	12.99	3.98	14.43
9	5.43	14.21	4.70	15.71
10	6.22	15.41	5.43	16.96

Here α is fraction outside the region of integration

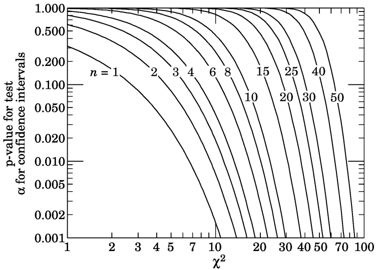
Introduction to Hypothesis Testing

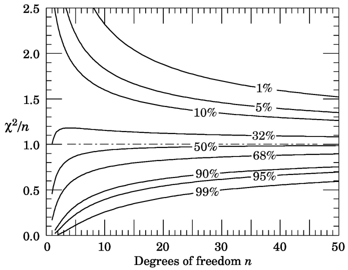
- So far, everything discussed geared to finding best value of parameters and uncertainy, under assumption that we know the pdf
- Nothing in our procedure tells us if data are consistent with hypothesis
- Need statistical tests of whether hypothesis is true
  - Significance tests: How likely is it that signal is just a fluctuation?
  - Goodness of fit tests: Is data consisten with coming from proposed hypothesis?
  - Exclusion tests: How big a signal could be hiding in our data?

Significance Tests

- Suppose we measure a value t for the data
  - How likely is it that we see a value that is further from prediction than our measurement
- Suppose we measure a distribution of data.
  - How consistent is our distribution with hypothesis
- Can use our friend χ^2

$$P\text{−}value = \int_{\chi^2_{meas}}^{\infty} f(x; n_d)dx$$





Hypothesis Testing: The Likelihood Ratio

- Experiments typically have background in addition to signal
- How do we know if there is a significant signal "on top of" the background?
- Given two hypotheses H_B and H_{S+B}, ratio of likelihoods is a useful test statistic

$$\lambda(\vec{N}) = \frac{\mathcal{L}(\vec{N}|H_{S+B})}{\mathcal{L}(\vec{N}|H_B)}$$